实验 2-3 kettle 实现源数据的预处理

建议课时：60 分钟

一、 实验目的

 熟练使用 kettle 工具做数据预处理；

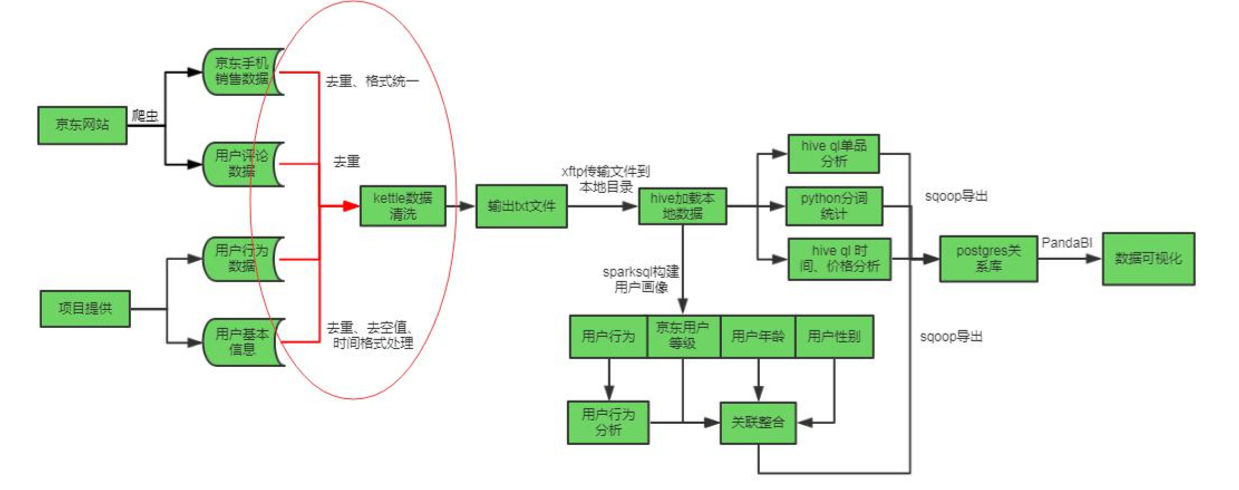
 掌握 kettle 中常用步骤的配置；

 能按清洗目标准确实现数据清洗工作；

二、 实验环境

Kettle7.0

三、 实验步骤

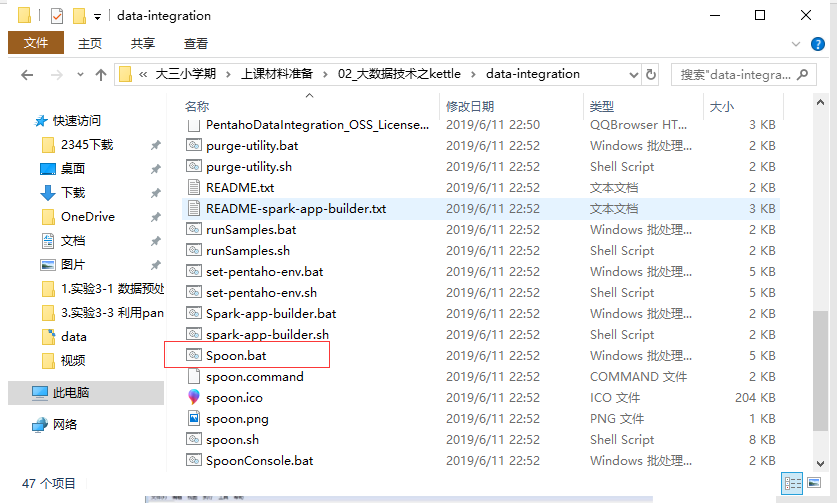


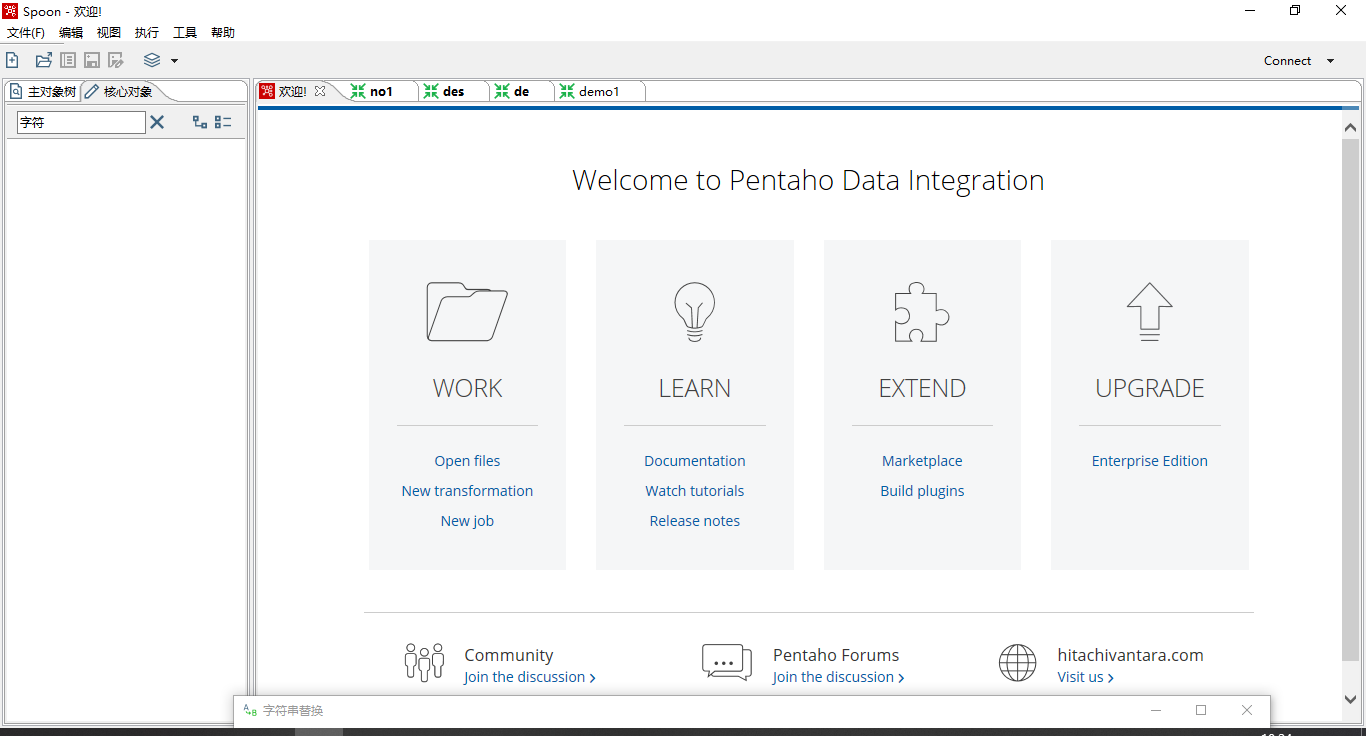
本节实验主要是通过 kettle 实现源数据的预处理。

具体实验步骤如下：

1. 下载安装 kettle7.0 工具

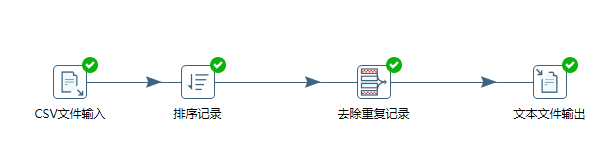
2. 运行 Spoon.bat 文件启动 kettle

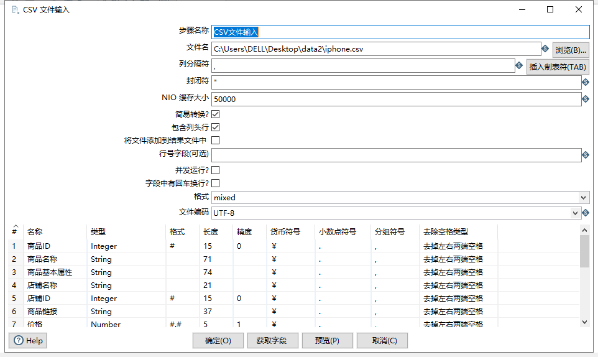


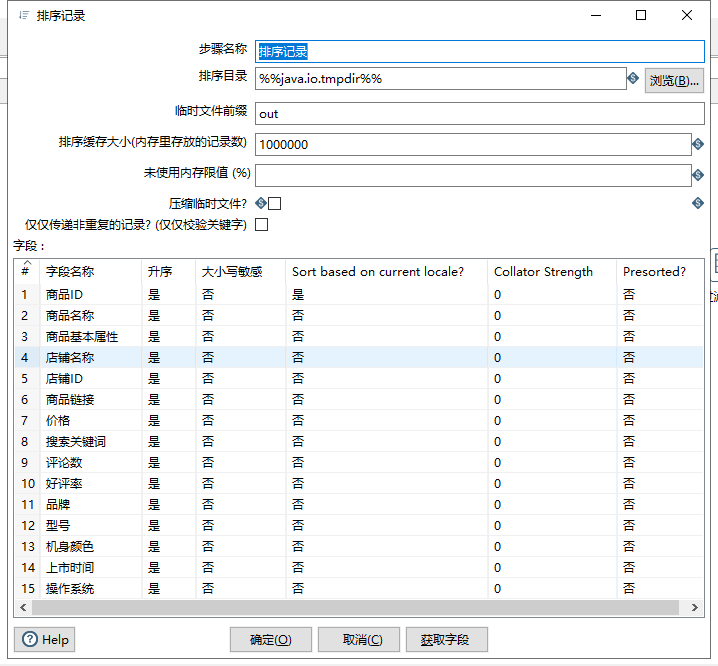


3. 新建转换去除手机销售信息表的重复记录

首先导入目标文件，之后进行排序，排序之后才能清洗（去重），之后导出文件：



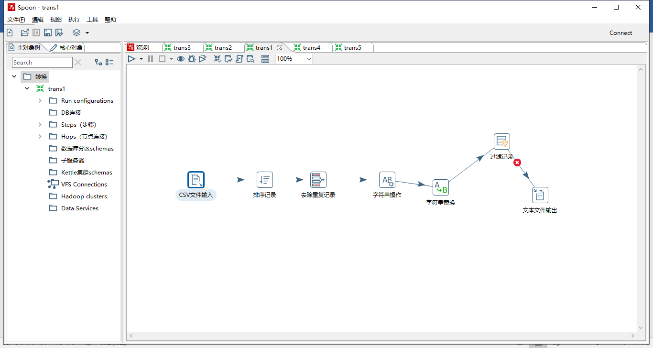


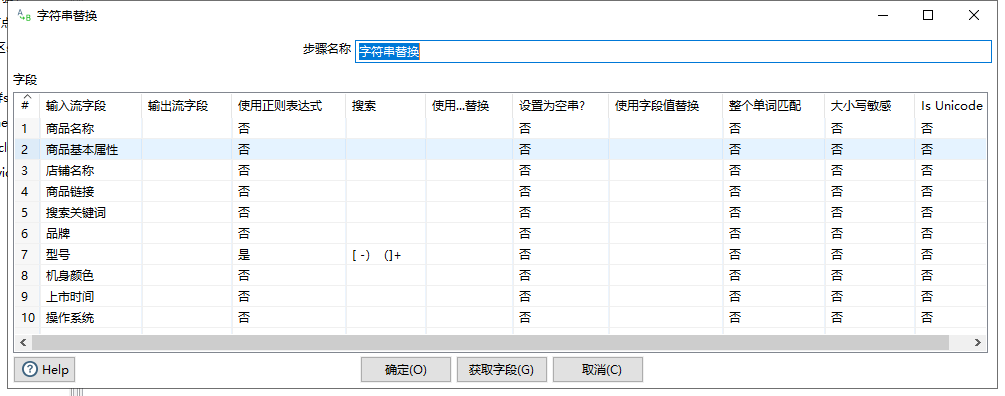


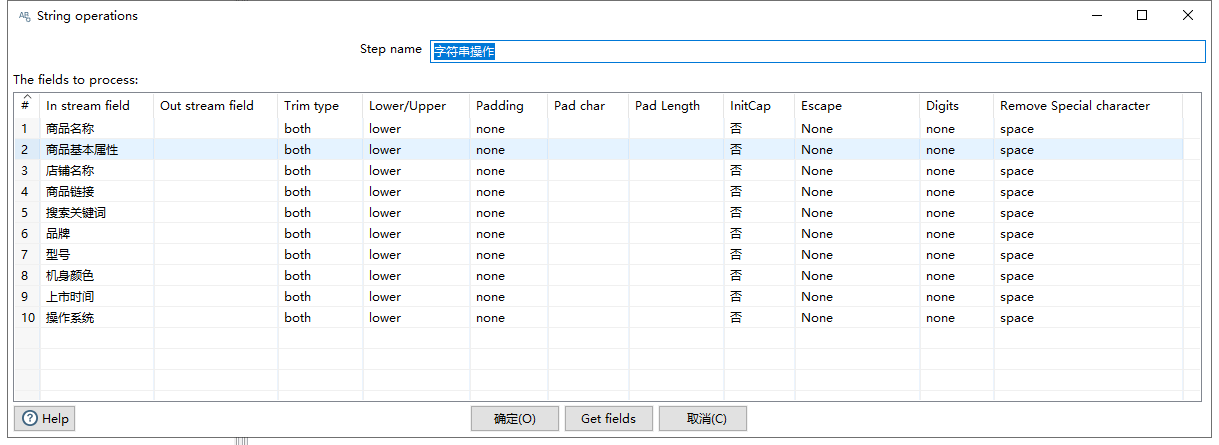


4. 在上述转换的基础上清洗手机销售信息表中型号字段中的数据

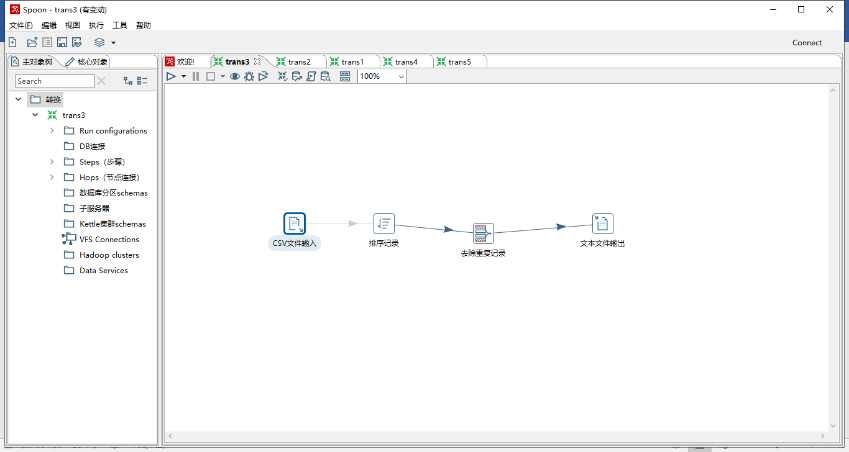
要求：去除该字段中的所有空格，方便后续聚合统计，字母统一大小写，去除该字段中的所有特殊字符（各种标点符号）



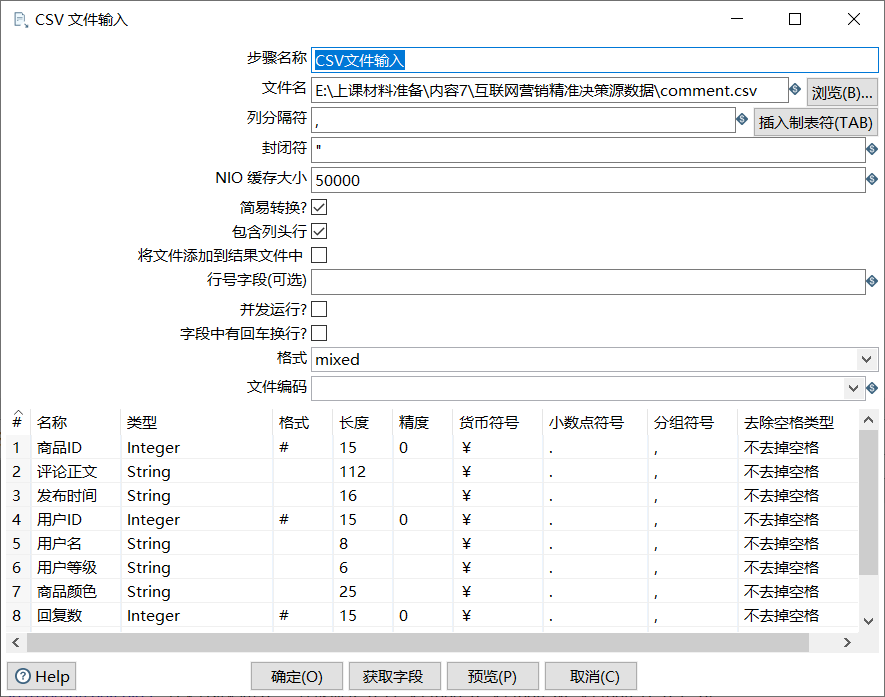


对于英文的品牌直接进行大小写统一（字符转换）：

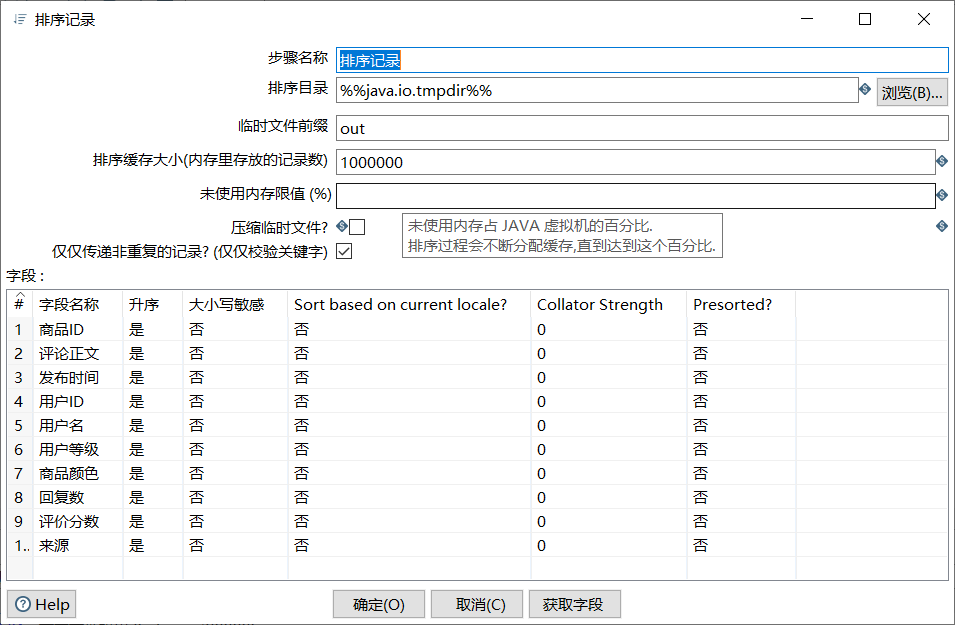
5. 新建转换去除用户评论信息表的重复记录



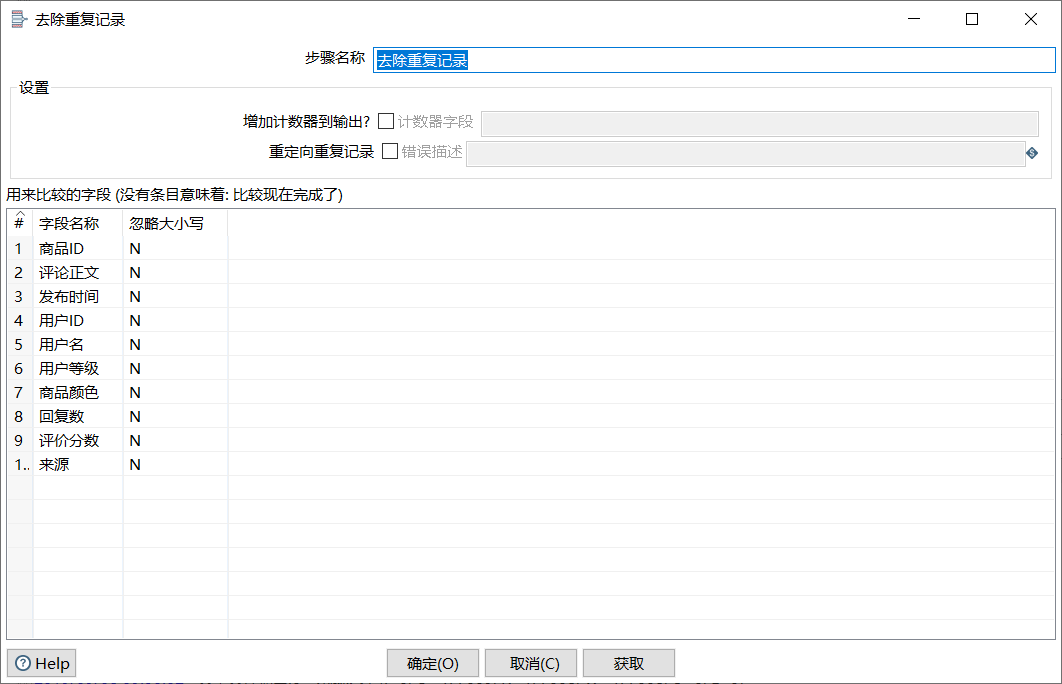
（2）在csv文件输入中选择要进行操作的.csv文件



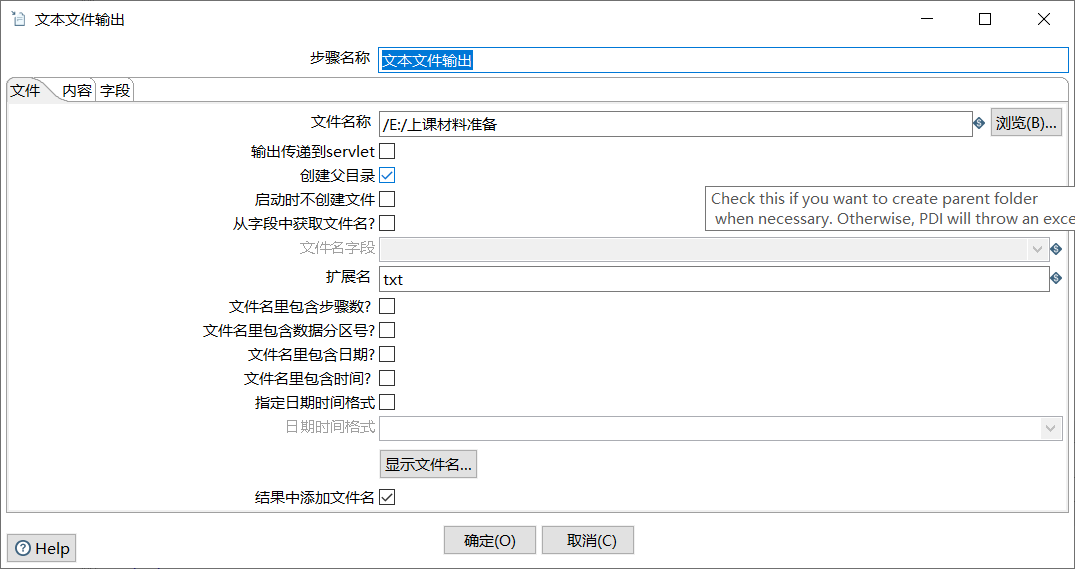
（3）在排序记录中获取字段



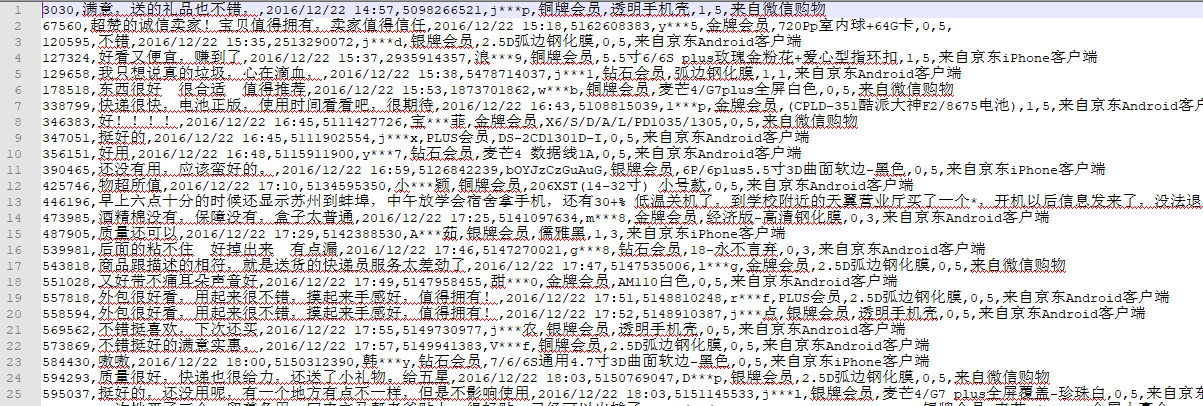
（4）在去除重复记录中获取字段



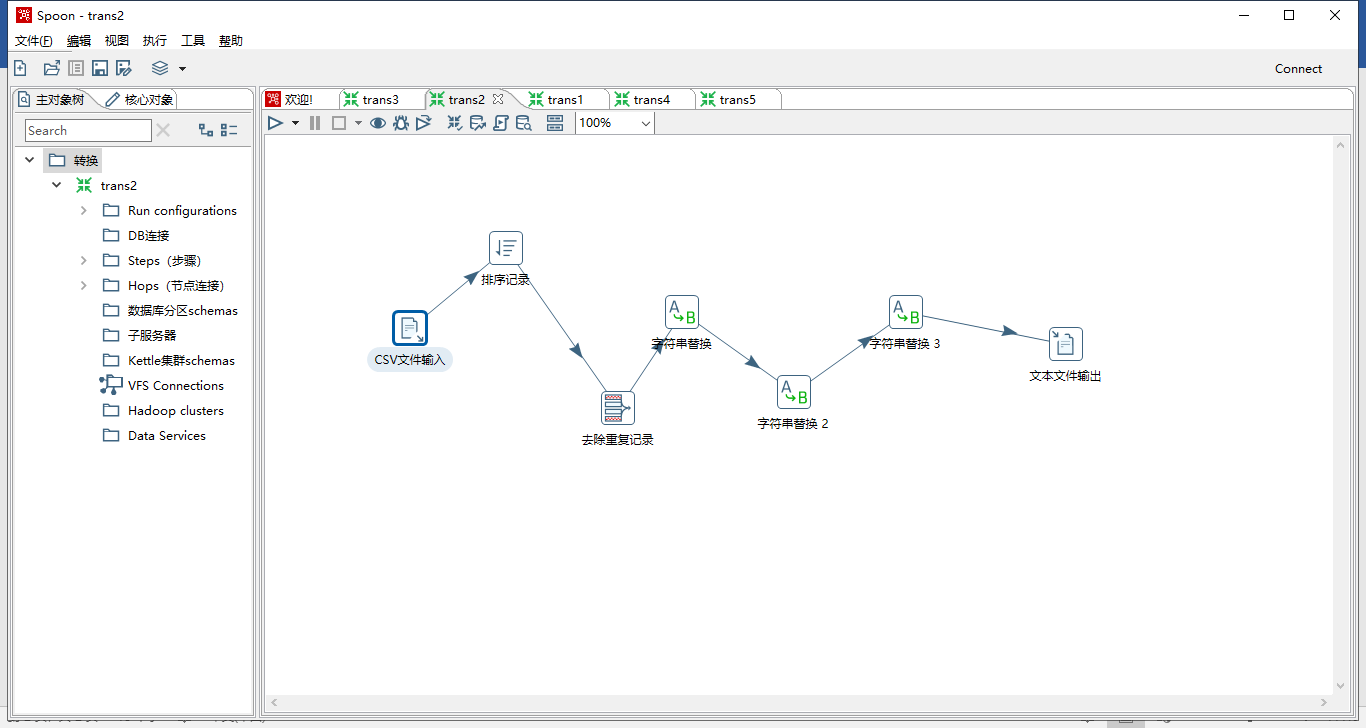
（5）在文本文件输出中选择输出的路径，进行输出



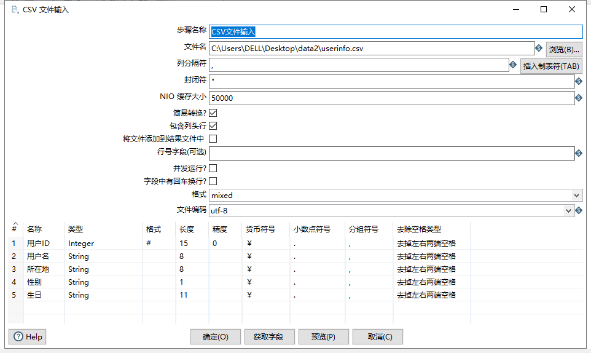
（6）生成的最终结果如下：



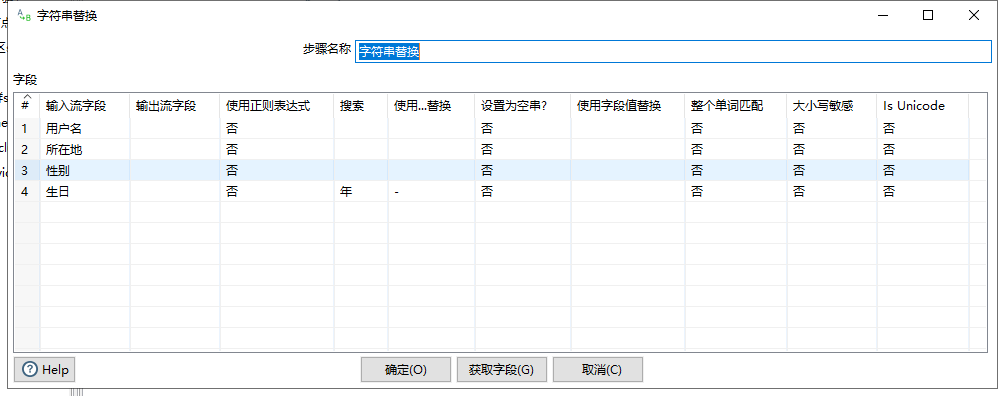
6. 新建转换处理用户信息表中出生日期字段（将 2019 年 5 月 20 日转换为2019-5-20）

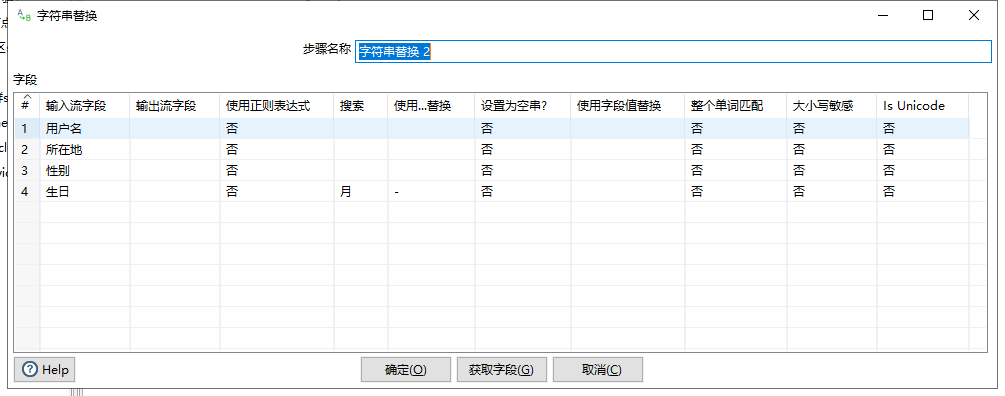


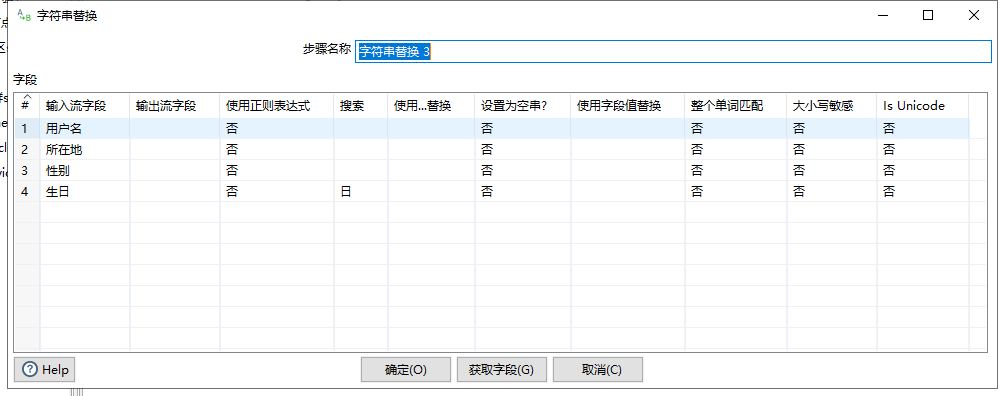
1. 在csv文件输入中选择要进行操作的.csv文件



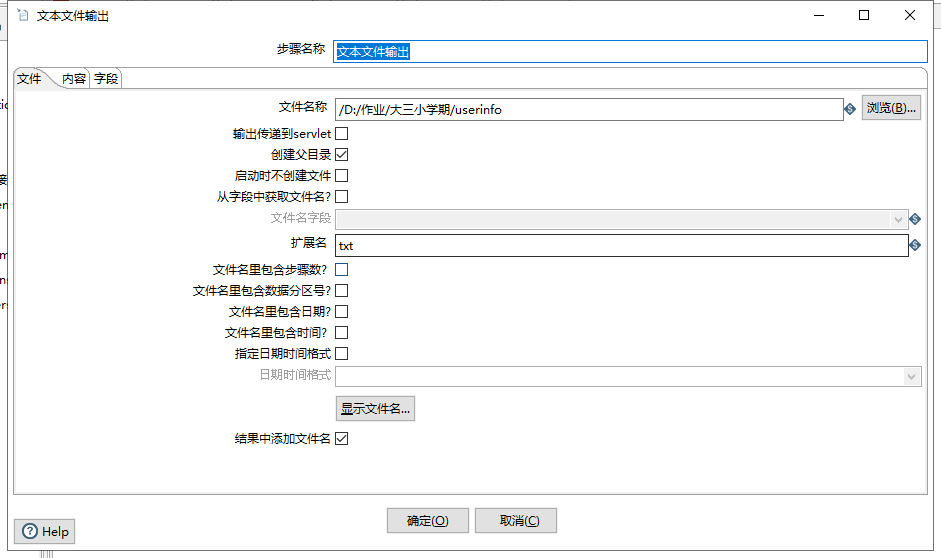
（3）建立3个替换将生日列中的年，月，日分别进行替换，将年，月替换成“-”，将日替换成空

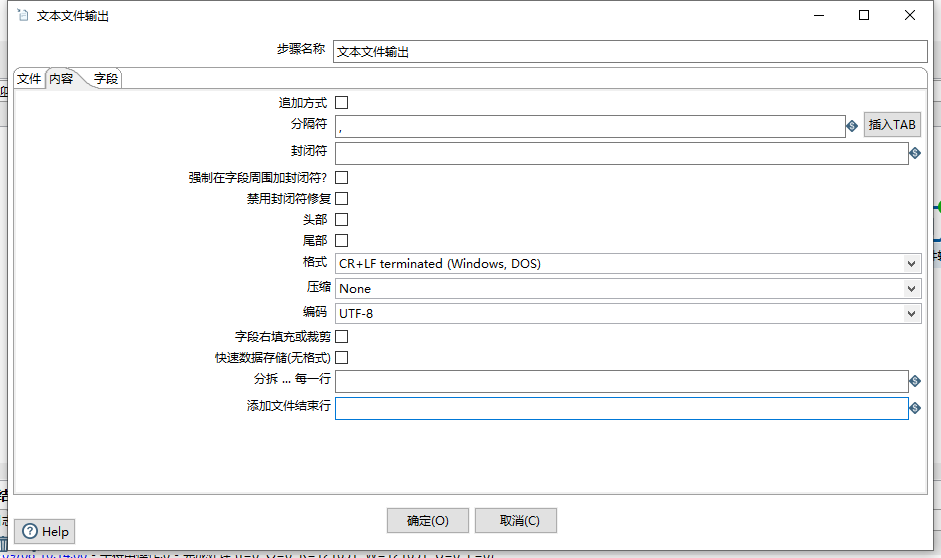




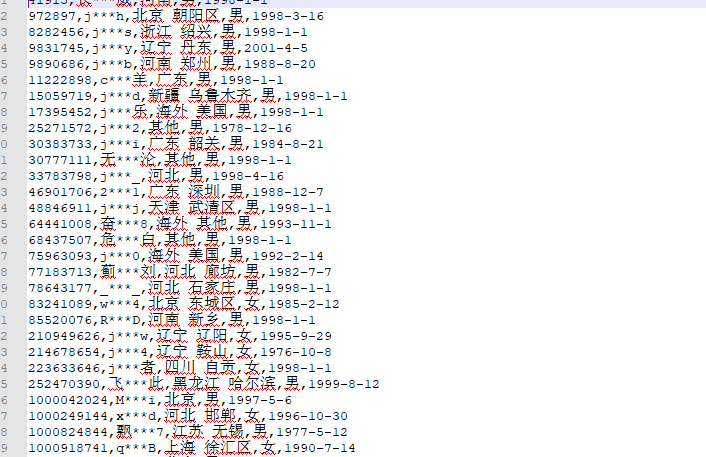


（4）在文本文件输出中选择输出的路径，进行输出





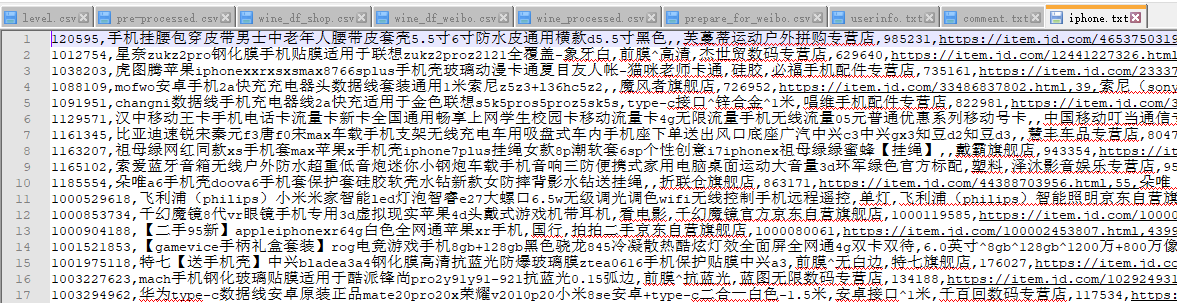
（5）输出结果如下：



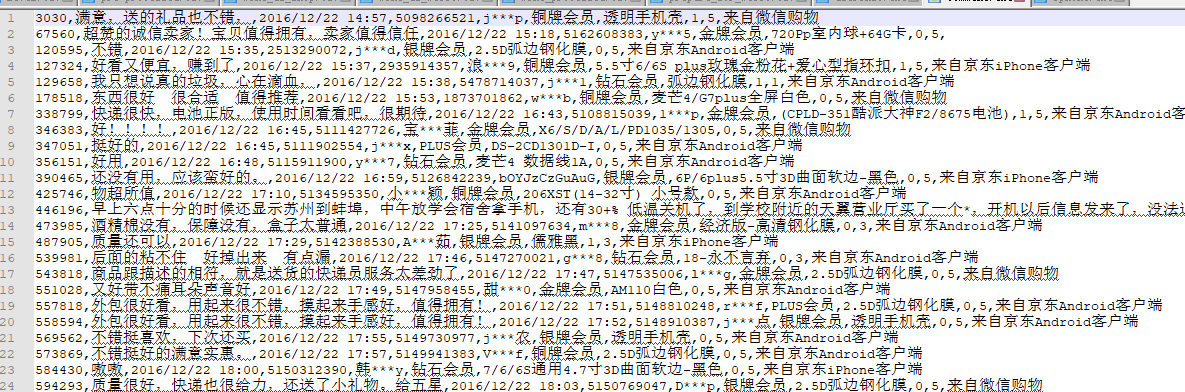
四、 实验成果

本次实验完成后，需要得到以下结果：

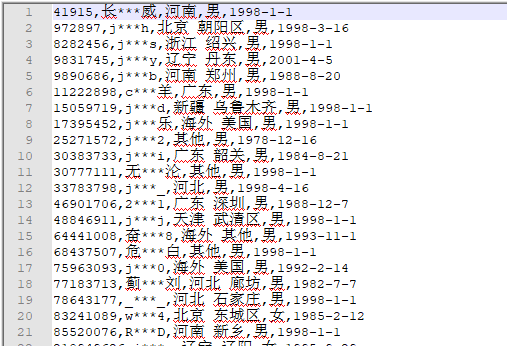
 去除手机销售信息表的重复记录；



 去除用户评论表的重复记录；



去除用户信息表的重复记录；年月份格式一致；



去除用户行为表的重复记录；年月份格式一致；

